Very Low Bitrate Spatial Audio Coding with Dimensionality Reduction

Christian Rohlfing, Jeremy E Cohen, Antoine Liutkus

▶ **To cite this version:**

Christian Rohlfing, Jeremy E Cohen, Antoine Liutkus. Very Low Bitrate Spatial Audio Coding with Dimensionality Reduction. 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar 2017, New Orleans, United States. hal-01515954

HAL Id: hal-01515954
https://inria.hal.science/hal-01515954

Submitted on 28 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

VERY LOW BITRATE SPATIAL AUDIO CODING WITH DIMENSIONALITY REDUCTION

Christian Rohlfing[1], *Jeremy E.Cohen*[2], *Antoine Liutkus*[3]

[1]Institut für Nachrichtentechnik, RWTH Aachen University, Germany
[2]Department of Image and Signal-processing, Gipsa-lab, CNRS, Grenoble, France
[3]Inria, speech processing team, Villers-les-Nancy, France

ABSTRACT

In this paper, we show that tensor compression techniques based on randomization and partial observations are very useful for spatial audio object coding. In this application, we aim at transmitting several audio signals called objects from a coder to a decoder. A common strategy is to transmit only the downmix of the objects along some small information permitting reconstruction at the decoder. In practice, this is done by transmitting compressed versions of the objects spectrograms and separating the mix with Wiener filters. Previous research used nonnegative tensor factorizations in this context, with bitrates as low as 1 kbps per object.

Building on recent advances on tensor compression, we show that the computation time for encoding can be extremely reduced. Then, we demonstrate how the mixture can be exploited at the decoder to avoid the transmission of many parameters, permitting bitrates as low as 0.1 kbps per object for comparable performance.

Index Terms— Spatial audio object coding, informed source separation, higher order SVD, dimensionality reduction

1. INTRODUCTION

Modern audio rendering technologies for the entertainment industry offer a high variability on the number of loudspeakers and their spatial configurations. In movie theaters for instance, offering an immersive audio user experience requires optimizing the rendering of each audio object depending on the geometry of the loudspeaker array [1]. Such applications motivated a lot of research activity in the topic of Spatial Audio Object Coding (SAOC) [2], which has its roots in Spatial Audio Coding (SAC) [3]. The core idea is to transmit many signals with only a downmix and upmixing parameters.

Similarly, *active listening* applications of musical content [4] include karaoke, sampling or individual stems equalization. They require interaction with the isolated sounds of a music mixture. In theory, this could be achieved by sound source separation techniques [5], which precisely aim at separating individual sounds from a mixture, see *e.g.* [6, 7, 8] for recent advances. However, the separation quality achieved there is often not sufficient for large audience applications. This fact motivated the incorporation of any available information about the signals to be recovered to increase performance, for instance midi scores, user annotations, etc. [9]. An interesting idea originally proposed in [10, 11] consists in using parameters for separation that were computed on the true sources beforehand during a *coding* stage. The originality of this approach is that those parameters are small enough to be conveyed along with the mixture in the metadata bitstream. This Informed Source Separation (ISS) setting leads to a substantial number of studies, all aimed at improving separation quality while reducing the size of the side information [12, 13, 14, 15].

This work is partly supported by ERC DECODA no. 320594, with european program FP7/2007-2013, and F.R.S.-FNRS incentive grant for scientific research no. F.4501.16

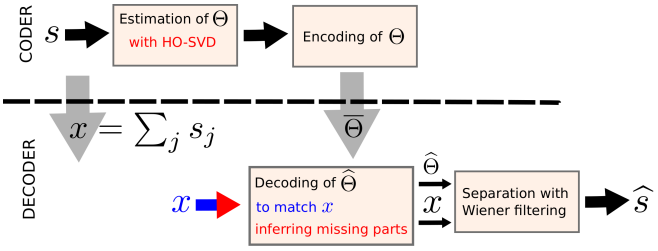

Fig. 1. General structure of a SAOC/ISS system. Black parts corresponds to the baseline methods [14], blue parts correspond to novelties brought in by [15, 16] and red parts to the current work.

As can be seen, ISS and SAOC are *in fine* different names for the same kind of architecture, depicted in Fig. 1. At the coding stage, J monophonic audio objects denoted s_j are available and used to compute side-information Θ. This side-information is quantized to yield $\overline{\Theta}$ and transmitted to the decoder, along with the downmix x of the objects. Then, at the decoder, the side-information is decoded as $\widehat{\Theta}$ and used to filter the mixtures in order to recover object estimates $\widehat{s}_j$. The remarkable feature of this framework is that it is much cheaper in terms of bitrate to transmit x and $\overline{\Theta}$ than all the objects s_j separately. Apart from the cost of transmitting x, recent studies such as [14] report 2 kbpso (kilobits per second per object) for $\overline{\Theta}$ with a reconstruction quality that is sufficient for active listening applications. For high fidelity applications, source coding strategies [17] report arbitrary distortions with minimal bitrate given by rate-distortion curves. To summarize, if a budget of 1-5 kbpso is available, very good reconstruction quality can be obtained at the decoder for SAC.

When very low bitrates are required, *i.e.* under 1 kbpso, classical methods such as [18, 14, 17] still fall short on providing good performance. In this setting, two options are indeed available. Either the number of parameters used for approximation of the sources are reduced, or they are very coarsely quantized. In either case, performance drops are too high for the methods to still be usable. To address this case, recent research [15, 16] depicted in blue on Fig. 1 exploit the computational power of the decoder to refine the quantized parameters $\overline{\Theta}$ to get closer to their original versions Θ. This is done by optimizing them again at the decoder to better explain the observed mixture. The rationale is that parameters which correctly describe the mixture should also correctly account for the sources, especially if they are initialized as $\overline{\Theta}$, *i.e.* close to the good solution Θ. This scheme allowed to reduce the bitrate to around 0.5 kbpso.

In this paper, we build on this previous research [15, 16]: we also pick a dimension reduction model for compressing the sources spectrogram and then we exploit the mixture at the decoding stage in order to refine the side-information. Our contributions in this respect are two-fold. First, we propose a computationally effective way to estimate the parameters Θ at the coder, that drops the nonnegativity constraint considered so far [14, 15, 16] and we rather make use of recent research on Higher-Order Singular Value Decompositions (HOSVD [19]). This allows to reduce the computational complexity of the estimation algorithm compared to previous research. Then, we propose not to send the whole HOSVD parameters to the decoder, but only part of them as degraded quantized values $\overline{\Theta}$, and we estimate them all again at the decoder using an iterative procedure. As we show, this strategy leads to transmission of side-information which size is independent of the track length, enabling bitrates as low as 0.1 kbpso, *i.e.* an order of magnitude less than the baseline [14] for similar performance.

2. TECHNICAL BACKGROUND

2.1. Notations

Let J be the number of audio signals s_j to transmit, that are called *sources* or *objects* equivalently. They are all monophonic and of the same length. In this paper, all processing is done in a Time-Frequency Representation (TFR) such as the short-term Fourier transform, which entries for object j are denoted $\boldsymbol{\mathcal{S}}(f,t,j) \in \mathbb{C}$. The resulting complex tensor $\boldsymbol{\mathcal{S}}$ has dimension $F \times T \times J$, where F is the number of non-redundant frequency bins and T is the number of time frames. We define the mix $x = \sum_j s_j$ as the monophonic sum of the objects. Its TFR is written $\boldsymbol{\mathcal{X}}$, with dimension $F \times T$ and entries $\boldsymbol{\mathcal{X}}(f,t) \in \mathbb{C}$. Now, with $\alpha \in (0,2]$, we define the *spectrograms* as

$$\boldsymbol{\mathcal{V}}_s(f,t,j) = |\boldsymbol{\mathcal{S}}(f,t,j)|^\alpha \quad \text{and} \quad \boldsymbol{\mathcal{V}}_x(f,t,j) = |\boldsymbol{\mathcal{X}}(f,t,j)|^\alpha,$$

and make the *additivity assumption* that the spectrogram of the mix can be approximated as the sum of the source spectrograms:

$$\boldsymbol{\mathcal{V}}_x(f,t) \approx \sum_j \boldsymbol{\mathcal{V}}_s(f,t,j). \tag{1}$$

Here, we take $\alpha = \frac{1}{2}$, although classical theory would suggest using power spectrograms $\alpha = 2$. It has indeed been shown experimentally (see e.g. [20]) that smaller α makes the assumption hold better. Now, if we assume the spectrograms of the sources are known, good estimates $\widehat{\boldsymbol{\mathcal{S}}}$ of the objects are given by Wiener filtering:

$$\widehat{\boldsymbol{\mathcal{S}}}(f,t,j) \leftarrow \mathbb{E}[\boldsymbol{\mathcal{S}} \mid \boldsymbol{\mathcal{V}}_s, \boldsymbol{\mathcal{X}}] = \frac{\boldsymbol{\mathcal{V}}_s(f,t,j)^{\frac{2}{\alpha}}}{\sum_j \boldsymbol{\mathcal{V}}_s(f,t,j)^{\frac{2}{\alpha}}} \boldsymbol{\mathcal{X}}(f,t)\ , \tag{2}$$

where the true spectrograms $\boldsymbol{\mathcal{V}}_s$ may be replaced by approximations $\widehat{\boldsymbol{\mathcal{V}}}_s$ in practice.

2.2. Informed source separation

To implement the architecture found in Fig. 1, most established techniques [14, 18, 2] use a Wiener filter (2) at the decoder, so that the parameters Θ are used to approximate the objects spectrograms $\boldsymbol{\mathcal{V}}_s$. Nonnegative Tensor Factorization (NTF [21]) is often used in this context. NTF approximates $F \times T \times J$ source spectrograms with as few parameters as $(F+T+J)R$, where R is called the number of components. Since R is typically very small compared to F and T, this leads to a tremendous reduction of the number of parameters to transmit.

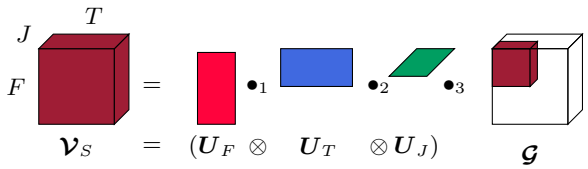

Fig. 2. HOSVD of tensor $\boldsymbol{\mathcal{V}}_s$. Product $\bullet_i$ is the contraction operator defined in Section 3.

Once the NTF parameters have been estimated at the coder based on the observation of the true $\boldsymbol{\mathcal{V}}_s$, they are quantized in the log-domain, as suggested by theory [17].

At the decoder, the classical approach [14, 18] implies using the quantized parameters directly to estimate the objects through Wiener filters. Alternatively, recent studies [15, 16] suggest taking the quantized values just as an initialization for fitting them again at the decoder, exploiting the relation (1) between mixture and sources spectrograms. The resulting parameters $\widehat{\Theta}$ are then used to filter the mix.

In any case, note that in the previous literature, *all parameters* are transmitted from the coder to the decoder, even if heavily quantized or re-estimated at the decoder. In this study, we propose an alternative model for spectrograms, as well as the option simply not to transmit the most costly parameters.

2.3. Higher-order SVD

Consider the source spectrogram tensor $\boldsymbol{\mathcal{V}}_s$, lying in a vector space $\mathbb{R}^{F\times T\times J}$. If it has been generated by a small number R of components conbined linearly, then $\boldsymbol{\mathcal{V}}_s$ follows a low rank model:

$$\boldsymbol{\mathcal{V}}_s = (\boldsymbol{F} \otimes \boldsymbol{T} \otimes \boldsymbol{J})\, \boldsymbol{\mathcal{I}}_R\ , \tag{3}$$

where $\boldsymbol{F} \in \mathbb{R}^{F\times R}$ collects the frequency signatures, $\boldsymbol{T} \in \mathbb{R}^{T\times R}$ the temporal signatures and $\boldsymbol{J} \in \mathbb{R}^{J\times R}$ their activation in each data set. $\boldsymbol{\mathcal{I}}_R$ is a diagonal tensor of size $R \times R \times R$ with ones on the diagonal. Here the operator $\otimes$ denotes a general tensor product, and $(\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{U}_J)$ is a multilinear mapping acting on tensors, *i.e.* linear with respect to each mode. In the two-way case, $(\boldsymbol{U}_F \otimes \boldsymbol{U}_T)\, \boldsymbol{V}_S = \boldsymbol{U}_F \boldsymbol{V}_S \boldsymbol{U}_T^\intercal$. Eq. (3) describes the well-known PARAFAC or Canonical Polyadic Decomposition model often used for blind source separation [22, 23, 24].

Typically R is much smaller than some dimensions of the tensor, so that $\boldsymbol{\mathcal{V}}_s$ actually lies in a low dimension subspace spanned on each mode by matrices $\boldsymbol{F}$, $\boldsymbol{T}$ and $\boldsymbol{J}$. In other words, if we choose orthogonal bases for each of the modes that span the three factor matrices of the PARAFAC model, $\boldsymbol{U}_F$, $\boldsymbol{U}_T$ and $\boldsymbol{U}_J$ of respective dimensions $F \times R_F$, $T \times R_T$ and $J \times R_J$, it can be shown that

$$\boldsymbol{\mathcal{V}}_s = (\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{U}_J)\, \boldsymbol{\mathcal{G}} \tag{4}$$

where $\boldsymbol{\mathcal{G}}$ is a small $R_F \times R_T \times R_J$ tensor of coefficients of $\boldsymbol{\mathcal{V}}_s$ in the orthogonal basis. If moreover some additional constraints on the slices of $\boldsymbol{\mathcal{G}}$ are satisfied, then this decomposition of $\boldsymbol{\mathcal{V}}_s$ is referred to as High Order Singular Value Decomposition (HOSVD) in the literature [19] and can be computed with high precision using SVDs of the tensor unfolded along each mode.

Since HOSVD maps a tensor $\boldsymbol{\mathcal{V}}_s$ to a smaller coefficient tensor $\boldsymbol{\mathcal{G}}$ in a feature space defined by three orthogonal matrices, it can be understood as a dimensionality reduction method for tensor data very similar in spirit to SVD for matrices, see Fig. 2.

3. DRISS: DIMENSION REDUCTION FOR ISS

3.1. Overview

In this paper, we propose the Dimension Reduction for ISS (DRISS) technique. It puts together ISS, presented in Section 2.2, and HOSVD, presented in Section 2.3, by replacing the classical NTF model of the baseline [14] by HOSVD (4). As we advocate, this leads to important computational savings at the coder as compared to the NTF approach.

Additionally, DRISS builds on previous research [15, 16] and involves computations at the decoder to refine the transmitted parameters $\overline{\Theta}$, that may be very coarsely quantized. The most remarkable feature of DRISS in this respect is that some elements of Θ are simply not included in $\overline{\Theta}$: those belonging to $\boldsymbol{U}_T$, pertaining to the temporal dimension. This brings a huge improvement in terms of bitrate for equivalent performance.

3.2. Coder-Decoder architecture

At the coder, the original data $\boldsymbol{\mathcal{V}}_s$ is available. The strategy of DRISS for obtaining a representation of $\boldsymbol{\mathcal{V}}_s$ with few parameters is to approximate $\boldsymbol{\mathcal{V}}_s$ with a structured tensor of low multilinear rank, so that:

$$\boldsymbol{\mathcal{V}}_s = (\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{U}_J)\, \boldsymbol{\mathcal{G}} + \boldsymbol{\mathcal{E}} \tag{5}$$

where $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{R_F \times R_T \times R_J}$ and $\boldsymbol{\mathcal{E}}$ stands for modeling error. The HOSVD parameters are then $\Theta = \{\boldsymbol{U}_F, \boldsymbol{U}_T, \boldsymbol{U}_J, \boldsymbol{\mathcal{G}}\}$.

It is easy to show that if $\boldsymbol{\mathcal{V}}_s$ is a low rank tensor, *i.e.* it can be described by a small number R of sources as in (3), then (5) is true with $\boldsymbol{\mathcal{E}} = 0$ as long as $R_F \geq \min(F, R)$, $R_T \geq \min(T, R)$ and $R_J \geq \min(J, R)$. Therefore, as long as the compression is not too large and the tensor actually follows a low rank model, $\boldsymbol{\mathcal{V}}_s$ lies exactly on a low dimension multilinear space and compression is not lossy. In practice, there is however a trade off between compressed dimensions and compression error.

As discussed in the previous section, one possible way to find a basis $(\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{U}_J)$ is to compute the HOSVD of $\boldsymbol{\mathcal{V}}_s$, also denoted as Maximum Likelihood SVD in the lossy case. Recall that HOSVD is well approximated by computing three SVDs of the three unfoldings of $\boldsymbol{\mathcal{V}}_s$. However, since the temporal dimension T can be very large, it is advantageous to resort to randomized version of the SVD as described in [25] for reducing both computation time and memory costs, yielding a very computationally effective coder. By using randomized methods, the numerical complexity of the HOSVD is then $\mathcal{O}\left(FT(R_F + R_T)\right)$ whereas NTF is typically computed by numerous iterations, each having numerical complexity $\mathcal{O}\left(FTJR\right)$.

Now, if the HOSVD is computed at the coder and Θ is transmitted as is, the number of parameters with high precision to be sent is $F \times R_F + T \times R_T + J \times R_J + R_F \times R_T \times R_J$. In the application at hand, the bottleneck is $\boldsymbol{U}_T$, which can be several orders of magnitude larger than the other parameters. For this reason, it is simply not included in the side-information $\overline{\Theta}$, while the others are heavily quantized.

At the decoder, we assume that $\boldsymbol{\mathcal{V}}_x$ is available, which obeys (1). Under the HOSVD model (4), it is hence possible to link $\boldsymbol{\mathcal{V}}_x$ and Θ as follows:

$$\boldsymbol{\mathcal{V}}_x \approx (\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{u}_J)\, \boldsymbol{\mathcal{G}}\,, \tag{6}$$

where $\boldsymbol{u}_J = \sum_{j=1}^{J} [\boldsymbol{U}_J]_{j:}$ is a row vector *i.e.* a linear form. This relation means that tensor $\boldsymbol{\mathcal{G}}$ is summed over the third mode using coefficients obtained by summing up the rows of $\boldsymbol{U}_J$, and transformed in the other two modes using exactly $\boldsymbol{U}_F$ and $\boldsymbol{U}_T$. Since the third mode of $\boldsymbol{\mathcal{V}}_x$ is contracted, *i.e.* since $\boldsymbol{\mathcal{V}}_x$ is a matrix, estimating $\boldsymbol{U}_J$ is not a feasible problem. Therefore, we simply assume that $\boldsymbol{U}_J$ is transmitted with high resolution, which is not a problem considering its very small dimension.

A straightforward strategy for estimating the parameters $\widehat{\Theta}$ of the low rank model (6) is to pick a least-squares solution and choose:

$$\widehat{\boldsymbol{U}}_F, \widehat{\boldsymbol{U}}_T, \widehat{\boldsymbol{\mathcal{G}}} \leftarrow \underset{\boldsymbol{U}_F, \boldsymbol{U}_T, \boldsymbol{\mathcal{G}}}{\operatorname{argmin}} \|\boldsymbol{\mathcal{V}}_x - (\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{u}_J)\, \boldsymbol{\mathcal{G}}\|_F^2 \tag{7}$$

with additional constraints $\widehat{\boldsymbol{U}}_F^{\mathsf{T}} \widehat{\boldsymbol{U}}_F = \boldsymbol{I}_F$ and $\widehat{\boldsymbol{U}}_T^{\mathsf{T}} \widehat{\boldsymbol{U}}_T = \boldsymbol{I}$. The Frobenius norm $\|\cdot\|_F^2$ is the sum of all squared coefficients.

Optimization problem (7) is non-convex and quite difficult to solve for both $\boldsymbol{U}_F$ and $\boldsymbol{U}_T$ simultaneously. However, we can choose an alternating procedure. Indeed, (6) is linear with respect to both $\boldsymbol{U}_F$ and $\boldsymbol{U}_T$, so that (7) becomes a simple singular value estimation problem with respect to these two matrices. Fixing $\boldsymbol{U}_F$ and $\boldsymbol{\mathcal{G}}$, (7) may be rewritten as:

$$\widehat{\boldsymbol{U}}_T \leftarrow \underset{\boldsymbol{U}_T^{\mathsf{T}} \boldsymbol{U}_T = \boldsymbol{I}}{\operatorname{argmin}} \|\boldsymbol{\mathcal{V}}_x - \boldsymbol{N}_T \boldsymbol{U}_T^{\mathsf{T}}\|_F^2 \tag{8}$$

where $\boldsymbol{N}_T = \boldsymbol{U}_F\, (\boldsymbol{u}_J \bullet_3 \boldsymbol{\mathcal{G}})$, with $\bullet_i$ standing for the contraction operator, *e.g.* on the first mode $[\boldsymbol{U} \bullet_1 \boldsymbol{\mathcal{V}}]_{ftj} = \sum_{k=1}^{F} \boldsymbol{U}_{fk} \boldsymbol{\mathcal{V}}_{ktj}$.

A well known solution to this problem is obtained by computing the SVD of $\boldsymbol{\mathcal{V}}_x^{\mathsf{T}} \boldsymbol{N}_T = \boldsymbol{A}_T \boldsymbol{\Sigma}_T \boldsymbol{B}_T^{\mathsf{T}}$ and set $\widehat{\boldsymbol{U}}_T = \boldsymbol{A}_T \boldsymbol{B}_T^{\mathsf{T}}$ where $\boldsymbol{\Sigma}_T$ is considered to be square (zeros are removed). This allows us to recover an estimate $\widehat{\boldsymbol{U}}_T$ as we decided not to transmit $\boldsymbol{U}_T$ at all.

However, we can even go further and try to re-estimate the quantized versions of the other parameters $\boldsymbol{U}_F$ and $\boldsymbol{\mathcal{G}}$. Similarly to (8), an update for $\boldsymbol{U}_F$ is obtained by setting $\boldsymbol{N}_F = (\boldsymbol{u}_J \bullet_3 \boldsymbol{\mathcal{G}})\, \boldsymbol{U}_F^{\mathsf{T}}$ and computing the SVD of $\boldsymbol{\mathcal{V}}_x \boldsymbol{N}_F^{\mathsf{T}}$.

Estimating $\boldsymbol{\mathcal{G}}$ when other parameters are known is not as straightforward. Indeed, since $(\boldsymbol{U}_F \otimes \boldsymbol{U}_T \otimes \boldsymbol{u}_J)$ is a linear operator on $\boldsymbol{\mathcal{G}}$, finding the best $\boldsymbol{\mathcal{G}}$ is equivalent to solving a linear system. However, the problem is ill-posed. Indeed, even though $\boldsymbol{U}_F$ and $\boldsymbol{U}_T$ admit a left inverse through the transposition operator, $\boldsymbol{u}_J$ is a linear form and therefore trivially does not admit a left inverse. This means that $\boldsymbol{\mathcal{G}}$ can be estimated by solely tensors $\widehat{\boldsymbol{\mathcal{G}}}$ having multilinear rank set to 1 on the third mode.

Clearly this is too restrictive since $\boldsymbol{\mathcal{G}}$ is not rank 1 on the third mode in the general case. To allow the estimate $\widehat{\boldsymbol{\mathcal{G}}}$ to be a full multilinear rank tensor all the same, an increment δ_G is introduced in (7). Again, finding the best $\delta_G = \widehat{\boldsymbol{\mathcal{G}}} - \boldsymbol{\mathcal{G}}$ means solving an underdetermined linear problem, and the least squares solution yields

$$\widehat{\delta_G} = \left(\boldsymbol{U}_F^{\mathsf{T}} \otimes \boldsymbol{U}_T^{\mathsf{T}} \otimes \boldsymbol{u}_J^{\mathsf{T}} \boldsymbol{u}_J / \|\boldsymbol{u}_J\|_F^2\right) \boldsymbol{\mathcal{V}}_x - \boldsymbol{u}_J \bullet_3 \boldsymbol{\mathcal{G}}\,. \tag{9}$$

Interestingly, $\widehat{\delta_G} = 0$ if $\boldsymbol{\mathcal{G}}$ is perfectly known.

In each of the updated rules described above, the cost function decreases, so that an alternating algorithm has to converge to some parameter set $\widehat{\Theta}$. However, since the cost function (7) is non-convex, the quality of the solutions obtained by the alternating algorithm highly depends on the initialization. Thus, the quality of the quantized parameters $\overline{\Theta}$ used to initialize the algorithm plays a critical role in the final estimate $\widehat{\Theta}$. The whole estimation procedure is summarized in Algorithm 1.

4. EXPERIMENTS

Since the size $\#\overline{\Theta} = F \times R_F + J \times R_J + R_F \times R_T \times R_J$ of the side-information $\overline{\Theta}$ of DRISS is independent of the number T of

Algorithm 1 Alternating algorithm for DRISS at the decoder

INPUTS: quantized parameters $\overline{\Theta}$, mixed data matrix $\boldsymbol{\mathcal{V}}_x$, number of iterations N_{it}.
Set $\boldsymbol{U}_F^{(0)} = \overline{\boldsymbol{U}}_F$, $\boldsymbol{\mathcal{G}}^{(0)} = \overline{\boldsymbol{\mathcal{G}}}$, $\boldsymbol{u}_J = \sum_{j=1}^{J} [\overline{\boldsymbol{U}}_J]_{j:}$
If $\overline{\boldsymbol{U}}_T$ not transmitted, initialize $\boldsymbol{U}_T^{(0)}$ with Eq. (8)
for i from 1 to N_{it} **do**
$\boldsymbol{U}_F^{(i)} = \boldsymbol{A}_F \boldsymbol{B}_F^{\mathsf{T}}$
where $\boldsymbol{A}_F \boldsymbol{\Sigma}_F \boldsymbol{B}_F^{\mathsf{T}} = \text{SVD}\left[\boldsymbol{\mathcal{V}}_x (\boldsymbol{u}_J \bullet_3 \boldsymbol{\mathcal{G}}^{(i-1)})^{\mathsf{T}} (\boldsymbol{U}_T^{(i-1)})^{\mathsf{T}}\right]$
$\widehat{\delta_G} = ((\boldsymbol{U}_F^{(i)})^{\mathsf{T}} \otimes (\boldsymbol{U}_T^{(i-1)})^{\mathsf{T}} \otimes \frac{\boldsymbol{u}_J^{\mathsf{T}} \boldsymbol{u}_J}{\|\boldsymbol{u}_J\|_F^2}) \boldsymbol{\mathcal{V}}_x - \boldsymbol{u}_J \bullet_3 \boldsymbol{\mathcal{G}}^{(i-1)}$
$\boldsymbol{\mathcal{G}}^{(i)} = \boldsymbol{\mathcal{G}}^{(i-1)} + \widehat{\delta_G}$
$\boldsymbol{U}_T^{(i)} = \boldsymbol{A}_T \boldsymbol{B}_T^{\mathsf{T}}$
where $\boldsymbol{A}_T \boldsymbol{\Sigma}_T \boldsymbol{B}_T^{\mathsf{T}} = \text{SVD}\left[\boldsymbol{\mathcal{V}}_x^{\mathsf{T}} \boldsymbol{U}_F^{(i)} (\boldsymbol{u}_J \bullet_3 \boldsymbol{\mathcal{G}}^{(i)})\right]$
end for
OUTPUTS: $\widehat{\Theta} = \left\{\boldsymbol{U}_F^{(N_{\text{it}})}, \boldsymbol{U}_T^{(N_{\text{it}})}, \boldsymbol{U}_J, \boldsymbol{\mathcal{G}}^{(N_{\text{it}})}\right\}$

time frames, a thorough realistic evalutation was conducted on full-length tracks, as opposed to what is usually done in the literature *e.g.* [13]. For this purpose, 10 full-length tracks taken from DSD100 database were considered, each consisting of $J = 4$ sources sampled at 44100 Hz: vocals, bass, drums and accompaniment. The track lengths vary from 2:20 to 4:50 minutes. The TFR consists of $F = 400$ Mel-filters computed from STFT of window size 93 ms with 50 % overlap. The resulting spectrograms then consist of $3000 \leq T \leq 6300$ frames.

The performance of separation is computed as the signal-to-distortion ratio (SDR, in dB) between the true and estimated sources in time domain. This SDR value is given in reference to the SDR obtained by an oracle estimator [26] which estimates optimal Wiener filter masks. This leads to a differential score δ_{SDR} that is averaged over the sources. The bitrate r is obtained with GZIP on $\overline{\Theta}$ as done in *e.g.* [18] and measured in kbps per object (kbpso).

For the above mentioned 10 songs, performance of DRISS was evaluated with $R_J = 4$ and all possible combinations of the following values of parameters:

- $R_F, R_T \in \{5, 10, 20, 60, 100\}$
- No quantization ($q = 0$) or quantization ($q = 1$) with $N_{\boldsymbol{\mathcal{G}}}, N_F \in \{5, 10, 50, 100, 300, 1000\}$ centroids for $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{U}_F$, respectively
- Number of iterations $N_{\text{it}} \in \{0, 10\}$.

This results in 24500 different experiments on full length tracks; each one leading to a rate-distortion value $(r, \delta\text{SDR})$-point. For each of the 4 different scenarios $q \in \{0, 1\}$, $N_{\text{it}} \in \{0, 10\}$, all corresponding points are filtered to yield their Pareto front per mixture and then averaged using the locally weighted scatter plot smoothing [27]. The resulting four rate-quality curves for the proposed method are displayed on Fig. 3a).

We also evaluated the performance of the NTF method [18] as well as plain HOSVD without quantization and with full transmission of all parameters of (4).

Several interesting facts are noticeable. First, NTF (blue curve with square-markers) outperforms HOSVD (solid blue curve) for small bitrates. The nonnegativity constraint leads to meaningful spectrograms estimates, while negative values are simply floored to 0 a posteriori with HOSVD, which is sub-optimal. Second, not sending $\boldsymbol{U}_T$ and using Eq. (8) for its estimation (solid yellow curve) leads to a significant decrease of bitrates (divided by 10) compared to HOSVD, still at q=0, and reaches comparable performance as NTF.

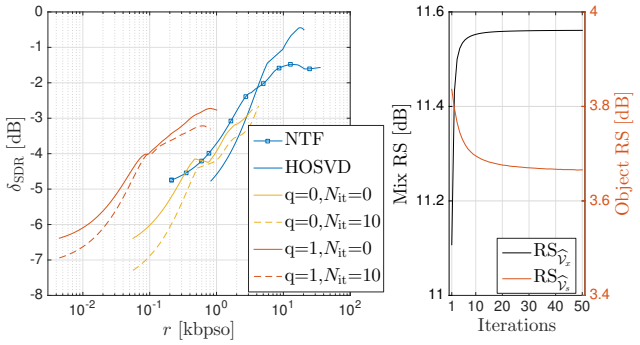

Fig. 3. Results: (a) rate-quality curves, (b) reconstruction scores

This means that reconstructing $\boldsymbol{U}_T$ at the decoder works. Third, including coarse quantization of the parameters ($q = 1$) leads to a remarkable further decrease of the bitrate by a factor of almost 10, for identical performance.

A startling fact however is that including several iterations in Algorithm 1 does not increase performance (dashed lines in Fig. 3a)). Investigating this fact, we calculated the average reconstruction scores for the mixture and the objects denoted with $\text{RS}_{\widehat{\boldsymbol{\mathcal{V}}}_x} = 10\log(\|\boldsymbol{\mathcal{V}}_x\|_F^2 / \|\boldsymbol{\mathcal{V}}_x - \widehat{\boldsymbol{\mathcal{V}}}_x\|_F^2)$ and $\text{RS}_{\widehat{\boldsymbol{\mathcal{V}}}_s}$ for each of $N_{\text{it}} = 50$ iterations of Algorithm 1. Fig. 3b) shows these scores for q=1 and the other parameters as chosen beforehand. Enabling the alternating algorithm does lead to a significant improvement of $\text{RS}_{\widehat{\boldsymbol{\mathcal{V}}}_x}$ (black curve) by 0.5 dB, proving that the algorithm does minimize the cost (7). However, the source score $\text{RS}_{\widehat{\boldsymbol{\mathcal{V}}}_s}$ (red curve) decreases slightly over iterations (about 0.2 dB). This shows again that iterations are not helping as already depicted in Fig. 3a). Just like in [15], we minimize a cost function at the decoder which leads to better describing $\boldsymbol{\mathcal{V}}_x$ and hopefully generalizing to $\boldsymbol{\mathcal{V}}_s$. Unlike in [15], this does not happen here. We explain this by the poor choice of a squared error used to model spectrograms, as well as by the lack of the nonnegativity constraint which may help [15] to generalize fitting $\boldsymbol{\mathcal{V}}_s$ from $\boldsymbol{\mathcal{V}}_x$. Future work may hence address these issues and focus on augmenting DRISS with constraints such as nonnegativity, different cost functions, or including additional helpful information to resolve the ambiguities raised by the estimation of $\boldsymbol{\mathcal{G}}$.

In any case, using HOSVD and estimating $\boldsymbol{U}_T$ with (8) at the decoder proves to be very effective as seen in Figure 3a). DRISS indeed yields similar performance than the NTF baseline, with a substantial decrease in both bitrate and computational cost, as discussed in Section 3.

5. CONCLUSIONS

In this paper, we proposed a dimension reduction technique for informed source separation (DRISS) that allows significant bitrate savings for spatial audio object coding compared to state of the art, with equivalent performance. Its two main ingredients are the following. First, it involves recently proposed randomized tensor compression techniques for fast encoding. Second, it avoids the transmission of all parameters thanks to an inference algorithm running at the decoder that is able to reconstruct the missing information by exploiting the available downmix. Interestingly enough, this approach also permits to recover from very coarse quantization of the model parameters.

As demonstrated in a thorough experimental study, DRISS leads to good reconstruction quality at very low bitrates around 0.1 kbps per object to encode, enabling wide range of applications.

6. REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio – the new standard for universal spatial/3D audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, 2015.

[2] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, et al., "Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[3] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial Audio Coding: Next-Generation Efficient and Compatible Coding of Multichannel Audio," in *Audio Engineering Society Convention 117*, Oct. 2004.

[4] C. Avendano and J-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[5] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*, Academic Press, 2010.

[6] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[7] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1562–1566.

[8] P. Smaragdis, C. Fevotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[9] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 257–260.

[10] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1464–1475, 2010.

[11] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721 –1733, Aug. 2011.

[12] S. Gorlow and S. Marchand, "Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2011, pp. 309 –312.

[13] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed source separation : a comparative study," in *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Aug. 2012.

[14] A. Liutkus, R. Badeau, and G. Richard, "Low bitrate informed source separation of realistic mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013.

[15] C. Rohlfing, J. Becker, and M. Wien, "NMF-based informed source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 474–478.

[16] C. Rohlfing and J. Becker, "Generalized constraints for NMF with application to informed source separation," in *2016 Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, Aug. 2016, pp. 597–601.

[17] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. on Audio, Speech and Language Processing*, 2012.

[18] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937 – 1949, 2012.

[19] L. De Lathauwer, B. De Moor, and Joos Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[20] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for nmf-based speech separation and music interpolation," in *IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2012, pp. 1–6.

[21] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley Publishing, Sept. 2009.

[22] F. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1, pp. 164–189, 1927.

[23] N. Sidiropoulos, L. De Lathauwer X. Fu, K. Huang, E. Papalexakis, and Christos C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *arXiv preprint arXiv:1607.01668*, 2016.

[24] J. E.Cohen, *Environmental multiway data mining*, Ph.D. thesis, Universite Grenoble Alpes, 2016.

[25] N. Halko, P. Martinsson, and J. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[26] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933 – 1950, Aug. 2007.

[27] W. Cleveland and S. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.